Cassandra Crossing/ Archivismi: archiviamo Cassandra, parte seconda

(565)—Dopo aver preparato i pdf non ci sono più scuse, dobbiamo archiviare il nostro primo articolo di Cassandra Crossing.

Cassandra Crossing/ Archivismi: archiviamo Cassandra, parte seconda
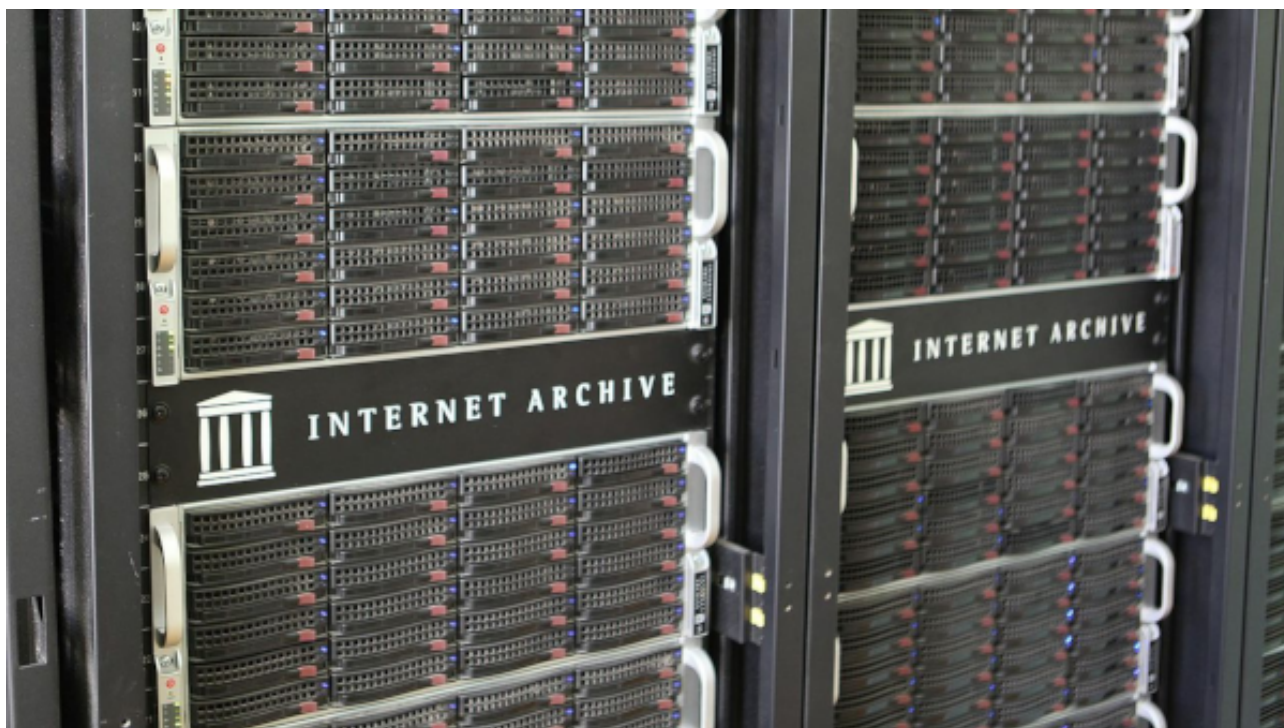
Figure 1:

(565)—Dopo aver preparato i pdf non ci sono più scuse, dobbiamo archiviare il nostro primo articolo di Cassandra Crossing.

1 gennaio 2024—Nelle precedenti puntate di *Archivismi* abbiamo raccontato le caratteristiche principali di Internet Archive, e caricato un semplice documento di esempio. Successivamente ci siamo dati l'ambizioso obbiettivo di uploadare l'*opera omnia* di Cassandra, ed abbiamo faticosamente preparato il materiale necessario nei formati e struttura più opportuni.

Non ci sono più scuse; è il momento di iniziare a caricare il primo documento di Cassandra Crossing, con tutte le cosette ed i metadati al posto giusto!

Dobbiamo quindi cimentarci davvero con *ia* e, visto che dovremo caricare centinaia di documenti, non farlo direttamente con la linea comandi, caricando un file per volta e scrivendo tutti i parametri ed i metadati su una lunghissima linea comandi.

Molto meglio impratichirsi fin da subito con i *bulk upload*, che si realizzano fornendo ad *ia* un unico parametro, cioè il nome di un foglio elettronico in formato CSV, in cui inseriremo i dati necessari (e li modificheremo tantissime volte per rimediare ad inevitabili errori).

Il comando per fare ciò è semplicemente

ia upload—spreadsheet=metadata.csv

Il lavoro vero sarà riempire il foglio elettronico finale con migliaia di righe di dati, ma facciamo un passo alla volta e carichiamo un solo oggetto, per cui un file di tre righe basterà.

Il nostro primo documento conterrà due file tra quelli generati per l'archiviazione, il *pdf* come documento principale e l'*html entrocontenuto* come secondo file; aggiungeremo anche un *minimo sindacale* di metadati, e l'identificativo verrà scelto uguale al nome dei file, tolta l'estensione.

Insomma, dopo molti, molti tentativi ecco il foglio ...

	A	B	C	D	E	F	G	H	I	J	K
1	identifier	file	description	subject[0]	subject[1]	subject[2]	title	creator	date	collection	mediatype
2	Test4_562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive	./html/562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive.html	Come archiviare gli articoli su Internet Archive	Soggetto 1	Soggetto 3	Soggetto 3	Archivismi: l'organizzazione e dei documenti in Internet Archive	Marco A.L. Calamari	2023	test_collection	texts
3	Test4_562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive	./pdf/562_Cassandra-Crossing--Archivismi--l-organizzazione-dei-documenti-in-Internet-Archive.pdf								test_collection	texts
4											

Figure 2:

Sembra facile, ma c'è voluta mezza giornata di lavoro, per avere il primo inserimento soddisfacente. Minuzie apparentemente insignificanti ma in realtà diaboliche hanno richiesto un sacco di tempo per prove e controprove. Ve ne racconto qualcuna qui, sperando così di farvi risparmiare tempo prezioso.

uno—quando salvate un foglio elettronico in formato CSV, che vuol dire "*valori separati da virgole*" non fidatevi della vostra applicazione. In certi casi, qui in Italia, l'applicazione potrebbe decidere di usare non la virgola ma il punto e virgola, e voi non ve ne accorgerete subito. Giuro, è successo!

due—disabilitate, nell'applicazione con cui state gestendo il foglio elettronico, tutti gli strumenti di autocorrezione; altrimenti il programma deciderà certamente di sostituire qualcosa per *il vostro bene*. Nel mio caso ha deciso di sostituire due segni meno consecutivi, presenti nei nomi di file, con un "*trattino lungo*", una modifica praticamente invisibile, anche da linea comandi. Questo ha portato all'inspiegabile messaggio di errore di *file non trovato*, ed ha rese necessarie alcune dozzine di prove, con relativi arrampicamenti sugli specchi. Non riferisco qui le parole che sono state pronunciate quando il problema è stato finalmente localizzato!

tre—state molto attenti quando inserite i valori nei campi. Un singolo spazio bianco prima o dopo il valore può non farlo interpretare, ed avere effetti imprevisti. Uno spazio all'inizio di " ***test_collection***" ha ad esempio impedito l'assegnazione corretta dell'oggetto alla *collection di test*, destinata, come già sapete, ad abilitare la cancellazione automatica dopo 30 giorni. In più considerate che non è possibile assegnare esplicitamente l'oggetto a collezioni pubbliche come "*opendata*", ma bisogna accettare la selezione automatica che verrà operata dal sistema.

quattro—inserite nel foglio la colonna *mediatype*, quando i documenti sono testuali (txt, html, pdf, etc.), ed usate il valore, "*texts*" altrimenti il sistema assegnerà automaticamente il valore "*data*" e questo avrà effetti collaterali insidiosi. Ad esempio il *browser di oggetti* non vi farà sfogliare le pagine, malgrado tutti i file derivati necessari siano stati creati correttamente. Il *me-*

diatype, contrariamente alla grande maggioranza dei parametri, non può più essere modificato, ma è necessario cancellare e rigenerare l'oggetto.

cinque—cancellare un oggetto non è un'operazione istantanea, ma richiede minuti o decine di minuti prima che l'effetto si propaghi in tutte la parti dell'interfaccia del sito. Non merita cancellare da linea comandi con *ia*; è decisamente più pratico farlo dalla pagina *My Upload*. Ricaricate spesso la pagina, e se notate cose strane, provate anche a svuotare la cache del browser.

sei—la comparsa di un oggetto appena creato nella finestra *My Upload* è, stranamente, abbastanza veloce, ma scatena tutte le operazioni "*derivative*", che a loro volta generano gli altri file in tempi variabili ma abbastanza lunghi. Questo vuol dire, ad esempio, che il *browser di oggetti* non sarà in grado di farvi sfogliare le pagine prima di una mezz'ora, e che la funzionalità di ricerca interna al *browser di oggetti* sarà attiva solo dopo parecchie ore.

Però, alla fine, che soddisfazione ...

Figure 3:

Ed anche per oggi è tutto. *Stay tuned* per la prossima puntata di "*Archivismi*".